Express Mail Label No. EV442788973US

[0001] DETERMINATION OF OPTIMAL SWITCHING POINTS BETWEEN THE UPLINK AND DOWNLINK

[0002] CROSS REFERENCE TO RELATED APPLICATION

[0003] This application claims priority from U.S. Provisional Patent Application Serial No. 60/457,941, filed March 26, 2003, which is incorporated by reference as if fully set forth herein.

[0004] FIELD OF THE INVENTION

[0005] This invention relates generally to wireless communication systems, and more particularly, to determining uplink and downlink resource allocations.

[0006] BACKGROUND

[0007] In many communication systems, uplink and downlink transmissions are separated, such as by frequency and/or time slots. One such system is the proposed wideband code division multiple access (WCDMA) frequency division duplex (FDD) mode, which separates the uplink and downlink by frequency. By contrast, the WCDMA time division duplex (TDD) mode separates the uplink and downlink by time slots, in response to uplink and downlink traffic demands.

[0008] For voice based communication systems, uplink and downlink demand is typically symmetrical, allowing for efficient symmetrical frequency allocations in FDD and time slot allocations in TDD type systems. Since more and more asymmetric wireless services are being utilized, such as Internet browsing, asymmetric allocations of frequencies/time slots are needed. To illustrate, in an FDD system, more downlink frequency bands may be needed than uplink or, in a TDD system, more downlink time slots may be needed than uplink. An incorrect allocation of these frequency bands/time slots leads to an under utilization of the radio resources.

[0009] Accordingly, it is desirable to have efficient approaches to allocating uplink and downlink resources.

[0010] SUMMARY

[0011] A wireless communication system has a variable number of time slots or frequencies allocated to support either uplink or downlink communications. Time slots or frequencies available for allocation to support either uplink or downlink communications are determined. Potential switching points between the available time slots or frequencies are determined. The switching points represent a change between time slots or frequencies used to support uplink and downlink communications. For each switching point, for each of uplink and downlink, a number of user that can be supported is determined by comparing a blocking probability of real time services with a required blocking probability of real time services and an average delay of non-real time services with a required average delay of non-real time services is compared. A minimum of the uplink and downlink users is selected that can be supported as the number of users that can be supported for that switching point. The switching point having a maximum number of users that can be supported is selected. The available uplink and downlink time slots or frequencies are allocated based on the selected switch point.

[0012] BRIEF DESCRIPTION OF THE DRAWING(S)

[0013] A more detailed understanding of the invention may be had from the following description of a preferred example, given by way of example and to be understood in conjunction with the accompanying drawing wherein:

[0014] Figure 1 illustrates state transitions of two adjacent states for real time services.

[0015] Figure 2 illustrates the state transition of two adjacent states for non-real time services.

[0016] Figure 3 is a flow diagram of frequency band/time slot switch point determination.

[0017] Figures 4A and 4B are simplified diagrams of a wireless system using optimum switching points.

[0018] DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0019] Although optimum switching point determination is described in conjunction with FDD and TDD wireless systems, such as W-CDMA FDD and TDD modes, time division synchronous CDMA (TD-SCDMA) and CDMA 2000, the embodiments are applicable to any communication system where the uplink and downlink are separated by variable resources. The following description is made in the context of a TDD system, although it can be applied to an FDD system by using frequency bands instead of time slots.

[0020] A TDD system supports both real-time and non-real-time services. If the number of wireless users, such as wireless transmit/receive units (WTRUs), in a cell is $N_{subscriber}$ and M-1 types of real-time service (j=1, 2, M-1), e.g., voice, video, etc., exist, each user of a real-time service has a Poisson call arrival rate of $\lambda_{subscriber}(j)$ for service type j. The total Poisson call arrival rate for service type j in the cell, denoted by λ_j, is $N_{subscriber} \cdot \lambda_{subscriber}(j)$.

[0021] The service time of service type j (j=1, 2, M-1) follows an exponential distribution with a mean of $1/\mu_j$. One type of generic non-real-time service is denoted by service type M. Each user has a Poisson packet arrival rate of $\lambda_{subscriber}(M)$ for the non-real-time service. The total Poisson packet arrival rate in the cell, denoted by λ_M, equals $N_{subscriber} \cdot \lambda_{subscriber}(M)$. The service time of the non-real-time packet follows an exponential distribution with a mean of $1/\mu_M$. The data rate of a service type j (j=1, 2, M) is R_j, and the required energy per bit to noise ratio is $(E_b/N_0)_j$. The requirement for

the blocking probability of real-time service type j (j=1, 2, ..., M-1) is assumed to be $P_{B\_req}(j)$, and the requirement for the average delay of non-real-time service is D_{req}.

[0022] Real-time calls are either admitted or blocked, while non-real-time packets can be buffered until resources are available. Therefore, real-time services have preemptive priority over the non-real-time service. For simplicity, the following assumes no priority exists between real-time services.

[0023] The load of an uplink time slot in TDD system is denoted by:

$$Load_{UL\_Slot} = (\beta_{UL} + \eta_{UL}) \cdot \sum_{i=1}^{N} \frac{1}{1 + \frac{W/S}{(E_b/N_0)_i \cdot R_i}} \qquad \text{Equation (1)}$$

β_{UL} is the multi-user detection (MUD) residual factor in the uplink, which is the fraction of intracell interference that cannot be cancelled by the MUD. η_{UL} is the average inter-to-intracell interference ratio in the uplink, and W/S is the equivalent chip rate of one time slot in a TDD system. When $Load_{UL\_Slot}$ approaches 1, the uplink capacity reaches its maximum, pole capacity, where the uplink interference goes to infinity.

[0024] Suppose that there are S_{UL} time slots in the uplink. ψ_j denotes $(\beta_{UL}+\eta_{UL}) \cdot \frac{1}{1+\frac{W/S}{(E_b/N_0)_j R_j}}$ for service type j. If there is a total of N_j users of service type j in the uplink, the total load of all uplink time slots can be expressed as:

$$Load_{UL} = \sum_{j=1}^{M} \psi_j \cdot N_j \qquad \text{Equation (2)}$$

[0025] Since the load of each uplink time slot has to be less than 1, the total load of S_{UL} time slots is less than S_{UL}.

[0026] The load of a downlink time slot in TDD system is expressed as:

[0027] $$Load_{DL\_Slot} = (\beta_{DL} + \eta_{DL}) \cdot \sum_{i=1}^{N} \frac{(E_b/N_0)_i \cdot R_i}{W/S} \qquad \text{Equation (3)}$$

β_{DL} is the MUD residual factor in the downlink, and η_{DL} is the average inter-to-intracell interference ratio in the downlink. When $Load_{DL\_Slot}$ approaches 1, the downlink capacity reaches its maximum, pole capacity, where the base station (BS) transmit power goes to infinity. Suppose that there are S_{DL} time slots in the downlink. ψ_j denotes $(\beta_{DL}+\eta_{DL})\cdot\frac{(E_b/N_0)_j \cdot R_j}{W/S}$ for service type j. If there are a total of N_j users of service type j in the downlink, the total load of all downlink time slots can be expressed as denoted in Equation (4) as:

$$Load_{DL} = \sum_{j=1}^{M} \psi_j \cdot N_j \qquad \text{Equation (4)}$$

[0028] Since the load of each downlink time slot is less than 1, the total load of S_{DL} time slots is less than S_{DL}.

[0029] In TDD systems, linearity between pole capacities of different servers may not exist as shown in equations (1) and (3). As a result, it cannot be modeled as a system that has a certain number of servers and wherein each user requests a certain number of servers.

[0030] The real-time services have preemptive priority over the non-real-time service. Therefore, the non-real-time services have no influence on the performance of real-time services. A multiple-class Markov chain is used to model the behavior of real-time services (j=1, 2, ..., M-1) in the system. For TDD systems, the load in the direction of interest (uplink or downlink) cannot exceed a certain maximum allowed value, denoted by $Load_{\max}$. $(X_1,...,X_{M-1})$ denotes the state where there are X_i calls of service type i in the system, and $P(X_1,...,X_{M-1})$ denotes the corresponding state probability. The allowed state for this system is denoted by Ω_{RT}, and is defined as:

$$\Omega_{RT} = \left\{ (X_1,...,X_{M-1}) \mid \sum_{j=1}^{M-1} \psi_j \cdot X_j \le Load_{\max} \right\} \qquad \text{Equation (5)}$$

[0031] Figure 1 shows the state transitions of two adjacent states. The flow balance equation is denoted as Equation (6) and, equivalently, by Equation (7):

$$\lambda_j \cdot P(X_1,...,X_j,...,X_{M-1}) = (X_j+1)\cdot\mu_j \cdot P(X_1,...,X_j+1,...,X_{M-1}) \quad \text{Equation (6)}$$

$$P(X_1,...,X_j+1,...,X_{M-1}) = \frac{\lambda_j}{\mu_j}\cdot\frac{1}{X_j+1}\cdot P(X_1,...,X_j,...,X_{M-1}) \quad \text{Equation (7)}$$

[0032] Using Equation (6) and Equation (7), Equation (8) and Equation (9) result:

$$P(X_1,...,X_{M-1}) = \left(\prod_{j=1}^{M-1}\left(\frac{\lambda_j}{\mu_j}\right)^{X_j}\cdot\frac{1}{X_j!}\right)\cdot P(0,...,0) \quad \text{Equation (8)}$$

$$\text{and} \sum_{(X_1,...,X_{M-1})\in\Omega_{RT}} P(X_1,...,X_{M-1}) = 1 \quad \text{Equation (9)}$$

[0033] The state probability $P(X_1,...,X_{M-1})$ is solved per:

[0034]
$$P(X_1,...,X_{M-1}) = \frac{\left(\prod_{j=1}^{M-1}\left(\frac{\lambda_j}{\mu_j}\right)^{X_j}\cdot\frac{1}{X_j!}\right)}{\sum_{(X_1,...,X_{M-1})\in\Omega_{RT}}\left(\prod_{j=1}^{M-1}\left(\frac{\lambda_j}{\mu_j}\right)^{X_j}\cdot\frac{1}{X_j!}\right)} \quad \text{Equation (10)}$$

[0035] The behavior of non-real-time services depends on how many real-time calls are being served in the system. Markov modulated Poisson process (MMPP) is used to model the behavior of non-real-time service in the system. $(X_M \mid X_1,...,X_{M-1})$ denotes the state when there are X_M non-real-time packets in the system given that there are X_i real-time calls of service type i in the system, and $P(X_M \mid X_1,...,X_{M-1})$ denotes the corresponding state probability. Since queuing is allowed for non-real-time services when all servers are busy, the allowed states for this system, Ω_{NRT}, becomes ∞. Non-real-time packets can only utilize the resources that are not used by real-time calls. The number of non-real-time packets that can be served when there are X_i real-time calls of service type i in the system is given by $\left\lfloor\left(Load_{\max} - \sum_{j=1}^{M-1}\psi_j \cdot X_j\right)/\psi_M\right\rfloor$, where $\lfloor x \rfloor$ is the largest integer that does not exceed x. With only X_M packets in the system, the actual throughput of non-real-time service (number of packets being served) is denoted by $T(X_M \mid X_1, X_2,...,X_{M-1})$ as:

[0036] $$T(X_M \mid X_1,\ldots,X_{M-1}) = min\left(X_M, \left\lfloor \left(Load_{\max} - \sum_{j=1}^{M-1} \psi_j \cdot X_j\right)/\psi_M \right\rfloor\right)$$ Equation (11)

[0037] The state transitions of two adjacent states are shown in Figure 2. The flow balance equation is denoted as:

[0038] $$\lambda_M \cdot P(X_M \mid X_1,\ldots,X_{M-1}) = T(X_M+1 \mid X_1,\ldots,X_{M-1}) \cdot \mu_M \cdot P(X_M+1 \mid X_1,\ldots,X_{M-1})$$ Equation (12)

[0039] Using Equation (12), Equation (13) and Equation (14) result:

$$P(X_M \mid X_1,\ldots,X_{M-1}) = \prod_{i=1}^{X_M}\left(\frac{\lambda_M}{\mu_M} \cdot \frac{1}{T(i \mid X_1,\ldots,X_{M-1})}\right) \cdot P(0 \mid X_1,\ldots,X_{M-1})$$ Equation (13)

and $$\sum_{X_M \in \Omega_{NRT}} P(X_M \mid X_1,\ldots,X_{M-1}) = 1$$ Equation (14)

[0040] $P(X_M \mid X_1,\ldots,X_{M-1})$ is solved per:

$$P(X_M \mid X_1,\ldots,X_{M-1}) = \frac{\prod_{i=1}^{X_M}\left(\frac{\lambda_M}{\mu_M} \cdot \frac{1}{T(i \mid X_1,\ldots,X_{M-1})}\right)}{\sum_{X_M \in \Omega_{NRT}} \prod_{i=1}^{X_M}\left(\frac{\lambda_M}{\mu_M} \cdot \frac{1}{T(i \mid X_1,\ldots,X_{M-1})}\right)}$$ Equation (15)

[0041] Since real-time services have preemptive priority over the non-real-time service, non-real-time service has no influence on the performance of real-time services. A service type i real-time new call will be blocked when the current load generated by real-time services plus the load of the new call exceeds the maximum allowed load.

[0042] B_i denotes the subset of states in which service type i new call will be blocked and is per:

[0043] $$B_i = \left\{(X_1, X_2,\ldots,X_{M-1}) \mid Load_{\max} - \psi_i < \sum_{j=1}^{M-1} \psi_j \cdot X_j \le Load_{\max}\right\}$$ Equation (16)

[0044] The blocking probability for service type i is given by the sum of state probabilities of those states that meet the blocking criteria.

[0045] $$P_{blocking}(i) = \sum_{(X_1,\ldots,X_{M-1}) \in B_i} P(X_1,\ldots,X_{M-1})$$ Equation (17)

[0046] The average number of non-real-time packets in the system, including packets waiting in the queue and packets being served, is denoted by $\bar{L}$ as follows:

[0047] $$\bar{L} = \sum_{(X_1,\ldots,X_{M-1})\in\Omega_{RT}} \left(\sum_{X_M\in\Omega_{NRT}} X_M \cdot P(X_M \mid X_1,\ldots,X_{M-1}) \right) \cdot P(X_1,\ldots,X_{M-1})$$ Equation (18)

[0048] The average throughput of non-real-time packets, X_i real-time calls of service type i in the system, is denoted by $\bar{T}_{|(X_1,\ldots,X_{M-1})}$ as follows:

$$\bar{T}_{|(X_1,\ldots,X_{M-1})} = \sum_{X_M\in\Omega_{NRT}} T(X_M \mid X_1,\ldots,X_{M-1}) \cdot P(X_M \mid X_1,\ldots,X_{M-1})$$ Equation (19)

[0049] The average throughput of non-real-time packets, denoted by $\bar{T}$ is per:

$$\bar{T} = \sum_{(X_1,\ldots,X_{M-1})\in\Omega_{RT}} \bar{T}_{|(X_1,\ldots,X_{M-1})} \cdot P(X_1,\ldots,X_{M-1})$$ Equation (20)

[0050] The average delay of non-real-time service is denoted by $\bar{D}$ as:

$$\bar{D} = \frac{\bar{L}}{\bar{T}} - \frac{1}{\mu_M}$$ Equation (21)

[0051] For $S_{dedicated}$ time slots used for dedicated physical channels, there are $S_{dedicated} - 1$ possible switching points. The switching point is the point where the resources are changed from uplink to downlink or vice versa. The number of feasible uplink (UL) time slots is S_{UL}, where S_{UL} =1, 2, ..., $S_{dedicated} - 1$, and the number of downlink time slots is $S_{DL} = S_{dedicated} - S_{UL}$. For each possible switching point, the number of users that can be supported in the uplink (denoted by $N_{\max\_UL}$) is determined as the largest number of users that satisfies the condition $P_{blocking}(j) \le P_{B\_req}(j),\ \forall j \in (1,2,\ldots,M-1)$ and $\bar{D} \le D_{req}$ in the uplink. Similarly, the number of users that can be supported in the downlink (denoted $N_{\max\_DL}$) is determined as the largest number of users that satisfies the condition $P_{blocking}(j) \le P_{B\_req}(j),\ \forall j \in (1,2,\ldots,M-1)$ and $\bar{D} \le D_{req}$ in the downlink. The number of users that can be supported for each switching point, denoted by $N_{\max}$, is given by $\min(N_{\max\_UL}, N_{\max\_DL})$. The switching point that yields the largest number of users that can be supported as the optimal switching point between uplink and downlink for the TDD system is selected.

[0052] Figure 3 is a flow diagram of optimum switch point determination for either a FDD or TDD system. The number of available frequency bands/time slots,

$S_{dedicated}$, is determined, (step 10). For each of the possible, $S_{dedicated}-1$, switching points, a maximum number of the users for the uplink and downlink is determined, (step 12). The maximum number of users is the number of users that have blocking probability of real-time services ($P_{blocking}(j)$) less than or equal to the required blocking probability ($P_{B\_req}(j)$) and have average delay of non-real time service ($\overline{D}$) less than or equal to the required delay (D_{req}). The users may be actual users of the system, if optimum switching is being used on a real time basis. As a result, the uplink and downlink resources can be dynamically changed. Alternately, the users may be based on statistical information on the service types typically used by the cell's users. As a result, the uplink and downlink resources may be fixed or changed periodically based on the statistical information.

[0053] For each switching point, the minimum number of users that can be supported in the uplink and downlink is selected as the number of users that can be supported by that switching point, (step 14). The switching point supporting the maximum number of users is selected, (step 16).

[0054] Figure 4A is a simplified diagram of an FDD system using optimum switching points. A radio network controller (RNC) 20 has a radio resource manager (RRM) 22. The RRM 22 determines a switching point (SP) between the available uplink and downlink frequencies. Frequencies to be used for the uplink and downlink are communicated to a Node-B 24. Although shown as a Node-B as for a third generation partnership (3GPP) communication system, the Node-B may be a base station, site controller, access point or other interfacing device in a wireless environment.

[0055] Uplink and downlink communications are transferred between the Node-B 24 and WTRUs 28_1 to 28_N (28) via an air interface 26. A WTRU includes but is not limited to a user equipment, mobile station, fixed or mobile subscriber unit, pager, or any other type of device capable of operating in a wireless environment. As illustrated, the air interface has P frequencies, F_1 to F_P, for the uplink, and S-P

frequencies for the downlink, F_{P+1} to F_S. As illustrated, the switching point (SP) is after P uplink frequencies out of the total of S available frequencies.

[0056] Figure 4B is a simplified diagram of a TDD system using optimum switching points. A RNC 20 has a RRM 22. The RRM 22 determines a switching point (SP) between the available uplink and downlink time slots. The time slots may be on one frequency band or multiple frequency bands. Time slots to be used for the uplink and downlink are communicated to the Node-B 24. Uplink and downlink communications are transferred between the Node-B 24 and WTRUs 28 via an air interface 26. As illustrated, the air interface has P time slots, TS_1 to TS_P, for the uplink, and S-P time slots for the downlink, TS_{P+1} to TS_S. The switching point (SP) is after P uplink time slots out of the total of S available time slots.

* * *